THE $\beta\alpha$-ENCODERS FOR ROBUST A/D CONVERSION

DAVID JIMÉNEZ AND YANG WANG

ABSTRACT. The β-encoder, introduced as an alternative to binary encoding in A/D conversion, creates a quantization scheme robust with respect to quantizer imperfections by the use of a β-expansion, where $1 < \beta < 2$. In this paper we introduce a more general encoder called the $\beta\alpha$-encoder, that can offer more flexibility in design and robustness without any significant drawback on the exponential rate of convergence of the obtained expansion. Mathematically, the $\beta\alpha$-encoder gives rise to a dynamical system that is both very interesting and challenging.

1. INTRODUCTION

Computer and digital information technologies are everywhere in our lives today. A key step that makes all those technologies possible is to convert analog data into digital ones, a process known *analog-to-digital conversion*, or A/D conversion. With we demand for higher precision and more cutting-edge technologies, the mathematics of A/D conversion algorithms plays a key role in this quest.

One of the most basic problems in A/D conversion concerns the representation of a signal x coming from a continuous media into a string of characters in a finite alphabet. Probably the best known scheme is the binary representation. In this scheme, a finite or infinite string of binary digits is obtained to represent $x \in [0, 1)$ in the following way

$$\begin{aligned} x_0 &= x \\ b_n &= Q(2x_{n-1}) \\ x_n &= 2x_{n-1} - b_n, \end{aligned} \tag{1.1}$$

where the function Q is given by

$$Q(t) = \begin{cases} 0 & \text{if } t < 1, \\ 1 & \text{otherwise.} \end{cases} \tag{1.2}$$

The function $Q(t)$ is called a *comparator* or a *quantizer*. Perfect reconstruction of x is given by

$$x = \sum_{n=1}^{\infty} b_n 2^{-n}. \tag{1.3}$$

The second author is partially supported by the the grant DMS-0813750 from National Science Foundation.

It is well know that the binary representation gives exponential accuracy in the sense that

$$\left| x - \sum_{n=1}^{N} b_n 2^{-n} \right| < 2^{-N}.$$

One important drawback of this A/D encoder is the fact that the binary representation is unique for all x that is not a dyadic rational, which has two binary representations. The significance of this drawback comes from the fact that in practice analog devices have inherent imprecisions. Thus if the comparator $Q(t)$ in (1.2) makes a wrong decision this error cannot be corrected. Thus the encoder based on binary expansion is inaccurate if the comparator $Q(t)$ is imprecise. Indeed, in practical applications it is more suited to model the comparator (quantizer) Q by the following Q_f with some build-in randomness:

$$Q_f(t) = \begin{cases} 0 & \text{if } t < \nu_1, \\ 0 \text{ or } 1 & \text{if } \nu_1 \le t \le \nu_2 \\ 1 & \text{otherwise,} \end{cases} \tag{1.4}$$

where the values of ν_1 and ν_2 are not known precisely, though, they lie in a known range. With such a "flaky" comparator the A/D encoder based on binary expansion will fail with high probability, see [2].

To address this concern the β-encoder for A/D conversion was recently introduced in [1] and studied in more detail in [2] and [4]. This encoder is based on the so-called β-expansion, which is analagous to the binary expansion but uses a value $1 < \beta \le 2$ in place of base 2. The β-encoder achieves exponential accuracy in the order of $O(\beta^{-N})$, but more importantly, with suitably chosen β the β-encoder is robust against imprecise comparators Q_f. In this paper we introduce a variant of the β-encoder called the $\beta\alpha$-encoder, where the introduction of a secondary parameter α allows more flexibility in the scheme while achieving the same exponential acuracy as in β-encoder.

2. The β-Encoder

The so called β-encoder is based in the β-expansion introduced originally in [10] as a particular case of an f-expansion. There, Renyi introduced the posibility to use non-integral bases to represent real numbers. Then, given a non integer $\beta > 1$, if $0 < x < 1$ it is possible to express

$$x = \sum_{n=1}^{\infty} b_n \beta^{-n}. \tag{2.1}$$

The *digits* b_n can be chosen recursively as

$$\begin{aligned} x_0 &= x, \\ b_n &= \lfloor \beta x_{n-1} \rfloor, \\ x_n &= \beta x_{n-1} - b_n, \end{aligned} \tag{2.2}$$

where $\lfloor \cdot \rfloor$ denotes the integer part. At each step, $0 \le b_i \le \lfloor \beta \rfloor$. There is an immediate gain using this representation instead of the representation obtained by an integral base: There are many possible choices of $\{b_n\}$ that still yield a valid reconstruction for x with the

expansion (2.1). In fact it is proved (see Sidrov [11]) that for almost every $x \in (0,1)$ there are uncontably many of such representations. Taking advantage of this fact, Daubechies, DeVore, Güntürk and Vaishampayan [2] introduced the idea of a β-encoder for A/D conversion, which enables one to overcome the imprecision of the comparator Q_f. (This is often referred to as the i.e. the *flaky quantizer* problem.) They proved the following theorem:

Theorem 1. *Let $1 < \beta < 2$, $0 \le x < 1$, $1 \le \nu_0 < \nu_1 \le (\beta-1)^{-1}$ and Q_f as defined in (1.4), and define x_n^f, b_n^f by the algorithm*

$$\begin{aligned} x_0^f &= x \\ b_n^f &= Q_f(\beta x_{n-1}^f) \\ x_n^f &= \beta x_{n-1}^f - b_n^f \end{aligned} \tag{2.3}$$

Then, for all $N \in \mathbb{N}$

$$0 \le x - \sum_{n=1}^{N} b_n^f \beta^{-n} \le \nu_1 \beta^{-N}.$$

Note that $\nu_1 \ge 1$. This means that even though the β-encoder allows certain imprecision on the quantizer, it does not allow the quantizer to err upward, i.e. reading off a 0 as a 1. The scheme would fail if this occurs. This drawback can be overcome by replacing $Q_f(t)$ with $Q_f(t-\delta)$ where δ is known to have $\delta \ge \nu_2$. This requires a conservative estimate of ν_1. In this paper we propose an alternative approach. We introduce the $\beta\alpha$-encoder as a variation of the β-encoder. With a secondary parameter α this encoder allows added flexibility.

3. The $\beta\alpha$-Encoder

As it has been already discussed, a β-expansion of a real number $x \in [0,1]$ is any collection of *digits* $\{b_n\}_{n\in\mathbb{N}}$ such that

$$x = \sum_{n\in\mathbb{N}} b_n \beta^{-n}.$$

Such expression is far from unique. A very intuitive way to obtain such a collection of digits is described by (2.2), and thus we will call this specific β-expansion of x as its *canonical expansion*. In this chapter we will analyze another way to obtain β-expansions, and will seize on the properties of this alternative method to obtain a stable scalar quantization scheme where the implementation can be given with some freedom unavailabe in the β-Encoder.

3.1. A Non-Canonical β-Expansion.

We will introduce a non-canonical β-expansion, that we will call a $\beta\alpha$-expansion. This one is similar to the β-expansion in that it still uses a possibly non-integer β as the base. However, unlike in the β-expansion the digits b_n are obtained at each stage using an amplification factor α instead of β. More precisely, for any $0 \le x < 1$ we set $x_0 = x$ and obtain b_n, x_n for $n \ge 1$ using the following scheme:

$$\begin{aligned} b_n &= \lfloor \alpha x_{n-1} \rfloor, \\ x_n &= \beta x_{n-1} - b_n. \end{aligned} \tag{3.1}$$

Observe that $x_{n-1} = \beta^{-1}(x_n + b_n)$ for every $n \geq 1$, and therefore, nesting this identity we obtain for any $N \in \mathbb{N}$ the expression

$$x = \beta^{-N} x_N + \sum_{n=1}^{N} b_n \beta^{-n},$$

or equivalently,

$$x - \sum_{n=1}^{N} b_n \beta^{-n} = \beta^{-N} x_N. \tag{3.2}$$

In order for perfect reconstruction $x = \sum_{n=1}^{\infty} b_n \beta^{-n}$ we will need $\beta^{-N} x_N \to 0$, preferably at an exponential rate. To make it happen, let $\{t\}$ denote the *fractional part* of t. Then $x = \lfloor x \rfloor + \{x\}$, and

$$\begin{aligned} x_N &= \beta x_{N-1} - b_N \\ &= \beta x_{N-1} - \lfloor \alpha x_{N-1} \rfloor \\ &= \beta x_{N-1} - \alpha x_{N-1} + \{\alpha x_{N-1}\} \\ &= (\beta - \alpha) x_{N-1} + \{\alpha x_{N-1}\} \\ &= (\beta - \alpha)^N x + \sum_{n=1}^{N} \{\alpha x_{n-1}\}. \end{aligned}$$

Since $0 \leq \{t\} < 1$, it follows that $(\beta - \alpha)^N x \leq x_N < (\beta - \alpha)^N x + N$, and

$$\beta^{-N}(\beta - \alpha)^N x \leq \beta^{-N} x_N < \beta^{-N}\big((\beta - \alpha)^N x + N\big).$$

Thus if we set $\beta > 1$ and $0 < \alpha \leq \beta$ we will ensure a perfect reconstruction with exponential rate convergence. Furthermore, all $x_n \geq 0$ and hence all digits b_n are nonnegative. For quantization applications, the magnitude of x_n matters because it determines the magnitude of b_n. Since these digits b_n must come from a finite alphabet we shall require that x_n be bounded. A necessary condition is $\beta - \alpha < 1$. In what follows we focus on the case $0 \leq \beta - \alpha < 1$. We ask the following questions: Are $\{b_n\}$ bounded, and if so, what is the upper bound?

Lemma 2. *Let $1 < \beta$, $\alpha \leq \beta$ and $\beta - \alpha < 1$. Define $T(x) = \beta x - \lfloor \alpha x \rfloor$ and set $\omega = [1 - (\beta - \alpha)]^{-1}$. Let $K = \lceil \omega(\beta - 1) \rceil$ where $\lceil y \rceil$ denotes the least integer greater than or equal to y. Then the fixed points of T are $\{k(\beta - 1)^{-1} : 0 \leq k < K\}$.*

Proof. First we notice that $T(x) \geq (\beta - \alpha)x$ implies that $T(x) > x$ if $x < 0$. So T cannot have a negative fixed point. Now, notice that if $T(x) = x$ then $\beta x - k = x$ where $k = \lfloor \alpha x \rfloor$. Thus $x = k(\beta - 1)^{-1}$. So all fixed points must be in the form of $x = k(\beta - 1)^{-1}$ for some integer $k \geq 0$. We shall determine which of these k's actually yield fixed points. To do so, let $x = k(\beta - 1)^{-1}$ be a fixed point. Then $\beta x - \lfloor \alpha x \rfloor = x$. It follows that $\lfloor \alpha x \rfloor = (\beta - 1)x = k$.

Now $\lfloor \alpha x \rfloor = \alpha x - \{\alpha x\}$. So we have $\alpha x - k = \{\alpha x\}$. Note that

$$\alpha x - k = \frac{\alpha k}{\beta - 1} - k = \frac{(1 - \beta + \alpha)k}{\beta - 1} = \frac{k}{\omega(\beta - 1)}.$$

Thus we have $k[\omega(\beta-1)]^{-1} = \{\alpha x\} < 1$, which yields $k < \omega(\beta-1)$ or equivalently, $k < K$. Conversely, if $0 \le k < K$ and $x = \frac{k}{\beta-1}$ the above calculations can be reversed to show that x is a fixed point. ■

Proposition 3. *Let $1 < \alpha \le \beta$ and $\beta - \alpha < 1$. Define $T(x) = \beta x - \lfloor \alpha x \rfloor$ and set*

$$M = \left\lceil \frac{\alpha(\beta-1)}{\beta[1-(\beta-\alpha)]} \right\rceil. \tag{3.3}$$

Let $\tau = M(\beta\alpha^{-1} - 1) + 1$. For any $0 \le x \le \tau$ we have $0 \le T^n(x) < \tau$ for all $n \ge 1$.

Proof. Note that $T(x) = (\beta-\alpha)x + \{\alpha x\}$ so $T(x) \ge 0$ for $x \ge 0$. Furthermore, as $\alpha < \beta$ we have $\alpha\beta^{-1}\omega(\beta-1) < \omega(\beta-1)$, where $\omega = [1-(\beta-\alpha)]^{-1}$. Thus $M \le \lceil \omega(\beta-1) \rceil$. Hence $(M-1)(\beta-1)^{-1}$ is a fixed point. For $x < M\alpha^{-1}$,

$$T(x) < (\beta-\alpha)x + 1 < (\beta-\alpha)M\alpha^{-1} + 1 = \tau.$$

If $M < \lceil \omega(\beta-1) \rceil$, then $M(\beta-1)^{-1}$ would be also a fixed point by Lemma 2 along with

$$\frac{\alpha(\beta-1)}{\beta[1-(\beta-\alpha)]} < M \Rightarrow M(\beta\alpha^{-1}-1) + 1 \le \frac{M}{\beta-1}.$$

Therefore, $T(x) \le x$ for every $x \in [M\alpha^{-1}, \tau)$. Hence $T(x) < \tau$.

If $M = \lceil \omega(\beta-1) \rceil$, then $(M-1)(\beta-1)^{-1}$ is the largest fixed point of T. Thus for every $x > M\alpha^{-1}$, $T(x) < x$.

As it was just proven, $0 \le x \le \tau$ implies $0 \le T(x) \le \tau$. The iteration step is trivial. ■

Proposition 4. *Let $1 < \alpha \le \beta$ and $\beta - \alpha < 1$. Let M and τ be as in Proposition 3. For any $x \in [0, \tau)$ define $x_0 = x$ and x_n, b_n for $n \ge 1$ by $b_n = \lfloor \alpha x_{n-1} \rfloor$ and $x_n = x_{n-1} - b_n$. Then $0 \le x_n < \tau$ and $b_n \in \{0, 1, \ldots, M\}$.*

Proof. Notice that $x_n = T^n(x_0)$. By Proposition 3 we have $0 < x_n < \tau$. Also, by $b_n = \lfloor \alpha x_{n-1} \rfloor$ it is enough to prove that $\tau\alpha \le M + 1$. Now, it follows from $\alpha < \beta$ that $\alpha(\beta-1) > \beta(\alpha-1)$. Thus

$$\frac{\alpha-1}{1-(\beta-\alpha)} < \frac{\alpha(\beta-1)}{\beta[1-(\beta-\alpha)]} \le M.$$

Hence $\alpha - 1 < M[1-(\beta-\alpha)]$, which yields $\tau = M(\beta-\alpha) + \alpha < (M+1)\alpha^{-1} \Rightarrow b_n \le M$. ■

3.2. **The $\beta\alpha$-Encoder vs. the β-Encoder.** The $\beta\alpha$-expansion described in the previous section leads naturally to a quantization scheme assuming a perfect quantizer. When a *flaky* quantizer is used, it can still yield a perfect reconstruction with suitable chioces of the parameters.

Given the conditions $1 < \alpha \le \beta$, $\beta - \alpha < 1$ we now consider the following general $\beta\alpha$-enocder given by the scheme $x_0 = x$,

$$\begin{aligned} b_n &= \bar{Q}(\alpha x_{n-1}), \\ x_n &= \beta x_{n-1} - b_n. \end{aligned} \tag{3.4}$$

where $\bar{Q}$ is a quantizer that such that $\bar{Q}(t) \in \{0, 1, \ldots, B-1\}$ for some integer B. The case $B > 2$ corresponds to a multi-bits quantizer, which is increasingly used in applications. Note that $\bar{Q}$ may have build-in uncertainty as in the 1-bit flasky comparator Q_f. Our main concern is to keep x_N bounded for every N.

The first natural question is: what bounds should x_N have to in order to have stability? Note that if $x_0 < 0$, then $x_1 = \beta x_0 - \bar{Q}(\alpha x_0) \leq \beta x_0$. Thus $x_N \leq \beta^{-N} x_0$, making the sequence diverge to negative infinity. Hence x_n should be positive. On the other hand, note that if x_N is bounded for all N, then by (3.2) one has that

$$x_N = \lim_{K\to\infty} \sum_{n=1}^{K} b_{N+n}\beta^{-n} \leq \sum_{n=1}^{\infty} (B-1)\beta^{-n} = \frac{B-1}{\beta-1}.$$

With $0 \leq x_n \leq \frac{B-1}{\beta-1}$ we have $x = \sum_{n=1}^{\infty} b_n \beta^{-n}$, and thus the exponential accuracy

$$0 \leq x - \sum_{n=1}^{N} b_n \beta^{-n} \leq \beta^{-N}. \tag{3.5}$$

The theorem below shows that even with an imprecise (multi-bits) quantizer $\bar{Q}$, which we shall denote by $\bar{Q}_f$, we can make $\{x_n\}$ bounded by suitably choosing the parameters β and α.

Theorem 5. *Let B be a given positive integer and let $\beta, \alpha > 1$ such that $1 < \beta < B$, $0 < \beta - \alpha < 1$. For any $x \in [0, 1)$ define $x_0^f = x$ and*

$$\begin{aligned} b_n^f &= \bar{Q}_f(\alpha x_{n-1}^f), \\ x_n^f &= \beta x_{n-1}^f - b_n^f, \end{aligned} \tag{3.6}$$

where the quantizer $\bar{Q}_f(t) \in \{0, 1, \ldots, B-1\}$ has the property $Q_f(t) = j$ implies that $t \in [j\alpha\beta^{-1}, \alpha\beta^{-1}(\mu + j)]$, with $\mu = (B-1)(\beta-1)^{-1}$. Then $0 \leq x_n^f \leq \mu$ for all n and hence

$$0 \leq x - \sum_{n=1}^{N} b_n^f \beta^{-n} \leq \mu\beta^{-N}.$$

Proof. Note that $\beta < B$ implies $\mu > 1$, and therefore $x_0^f < \mu$. Now $x_n^f = \beta x_{n-1}^f - b_n^f$ and (3.2) is valid regardless of how b_n^f are chosen. Therefore it suffices to prove that $0 \leq x_n^f \leq \mu$. Let's now examine the respective subintervals.

Assume that $0 \leq x_n^f \leq mu$. If $\bar{Q}_f(\alpha x_n^f) = j$ then $\alpha x_n^f \in [j\alpha\beta^{-1}, \alpha\beta^{-1}(\mu + j)]$. Thus $j\beta^{-1} \leq x_n^f \leq \beta^{-1}(\mu + j)$. It follows that $0 = \beta(j\beta^{-1}) - j \leq x_{n+1}^f \leq \beta[\beta^{-1}(\mu + j)] - j = \mu$. By induction on n we have $0 \leq x_n^f \leq mu$ for all n. ∎

By Proposition 4 we may choose $B = M + 1$ where $M = \left\lceil \frac{\alpha(\beta-1)}{\beta[1-(\beta-\alpha)]} \right\rceil$. This $\beta\alpha$-encoder gives a robust multi-bits encoder. A special case is to choose β and α so $B = 2$, which leads to a robust 1-bit $\beta\alpha$-encoder. The following theorem is a corollary of Theor.me 5, which is a generalization of the 1-bit β-encoder.

Theorem 6. *Let $1 < \beta < 2$ and $\alpha > 1$. Let Q_f be as in (1.4. Assume that $\beta(\beta-1) < \alpha < \beta$ and $\alpha\beta^{-1} \le \nu_1 < \nu_2 \le \alpha\beta^{-1}(\beta-1)^{-1}$ and). Then for any $x \in [0,1)$ the encoder given by*

$$\begin{array}{rcl} x_0^f &=& x, \\ b_n^f &=& Q_f(\alpha x_{n-1}^f), \\ x_n^f &=& \beta x_{n-1}^f - b_n^f \end{array} \tag{3.7}$$

has the property that for all $N \in \mathbb{N}$

$$0 \le x - \sum_{n=1}^{N} b_n^f \beta^{-n} \le (\beta-1)^{-1}\beta^{-N}.$$

3.3. **Imprecise α-Multiplication.** An imprecise quantizer is not the only problem that can arise in a real application. The multiplication via analog circuits can potentially be another source of inaccuracy. Thus, by performing two multiplications in the $\beta\alpha$-encoder we introduce an extra source for potential errors. In this section, we show that the α-multiplication in the $\beta\alpha$-encoder does not have to be very accurate, by allowing each α-multiplier to multiply a different value each time. We prove the following theorem.

Theorem 7. *Let B be a given positive integer and let $\beta, \alpha_n > 1$ such that $1 < \beta < B$, $\frac{\beta}{B} < \beta - \alpha_n < 1$. For any $x \in [0,1)$ let $x_0^f = x$ and*

$$\begin{array}{rcl} b_n^f &=& \bar{Q}_f(\alpha_n x_{n-1}^f), \\ x_n^f &=& \beta x_{n-1}^f - b_n^f, \end{array}$$

where the quantizer $\bar{Q}_f(t) \in \{0, 1, \ldots, B-1\}$ has the property $Q_f(t) = j$ implies that $t \in [j(\sup\alpha_k)\beta^{-1}, (\inf\alpha_k)\beta^{-1}(\mu + j)]$ and $Q_f(t) = B-1$ if $t \ge (\inf\alpha_k)\mu$, with $\mu = (B-1)(\beta-1)^{-1}$. Then, $0 \le x_n^f \le \mu$ for all n and hence

$$0 \le x - \sum_{n=1}^{N} b_n^f \beta^{-n} \le \mu\beta^{-N}.$$

Proof. Note that trivally, for any integer n and $0 \le j < B$ one has that

$$[j(\sup\alpha_k)\beta^{-1}, (\inf\alpha_k)\beta^{-1}(\mu + j)] \subseteq [j\alpha_n\beta^{-1}, \alpha_n\beta^{-1}(\mu + j)].$$

Then, using the same argument as in the proof of Theorem 5 we only need to prove the set of intervals $I_j = [j(\sup\alpha_k), (\inf\alpha_k)(\mu + j)]$ cover $[0, (\inf\alpha_k)\mu\beta]$. Note that $0 \in I_0$ and $(\inf\alpha_k)\mu \in I_{B-1}$, and the lower endpoints (as well as the upper endpoints) are in an increasing order. Thus the only thing left to prove is that $I_j \cap I_{j+1} \ne \emptyset$. For this it suffices to prove $(j+1)\sup\alpha_k \le (\mu + j)\inf\alpha_k$. Note that by assumption we have

$$\frac{\mu + j}{j+1}\inf\alpha_k \ge \frac{\mu + (B-1)}{B}(\beta - 1) = \frac{(B-1)\beta}{B} \ge \sup\alpha_k.$$

■

4. Ergodic Properties of the $\beta\alpha$-Encoder

In the previous chapter we discussed the $\beta\alpha$-Encoder. The scheme defined in (3.1) gives rise to the dynamical system $x_{n+1} = T(x_n)$, where $T(x) = \beta x - \lfloor \alpha x \rfloor$. Beyond its practical applications, this system is interesting on its own from a mathematical point of view, specifically, the ergodicity of T, on which we will focus in this section.

4.1. **Invariant Sets for T.** Let $1 < \beta$, $\alpha \leq \beta$ and $\beta - \alpha < 1$. As in Lemma 2, denote $T(x) = \beta x - \lfloor \alpha x \rfloor$, $\omega = [1 - (\beta - \alpha)]^{-1}$and $K = \lceil \omega(\beta - 1) \rceil$.

For simplicity we will introduce the following additional notation. For $0 < k \leq K$ let

$$\lambda_k = \frac{k}{(\beta - 1)}, \quad \xi_k = k\left(\frac{\beta - \alpha}{\alpha}\right) + 1, \quad \zeta_k = k\left(\frac{\beta - \alpha}{\alpha}\right). \tag{4.1}$$

By Lemma 2, λ_k are the nonzero fixed points of T. Observe that the map $T(x) = (\beta - \alpha)x + \{\alpha x\}$ is piecewise linear with discontinuities at $y_k = k/alpha$. It is easy to see that $\xi_k = \lim_{x \to y_k^-} T(x)$ and $\zeta_k = \lim_{x \to y_k^+} T(x)$.

Proposition 8. *If i and j are indices such that $\lambda_{i-1} \leq \zeta_i$ and $\xi_j \leq \lambda_j$, then $\zeta_i < \xi_j$. Furthermore the interval $\Psi = [\zeta_i, \xi_j]$ is T-invariant in the sense that $\overline{T(\Psi)} = \Psi$.*

Proof. Note that

$$\begin{aligned} \lambda_{i-1} &\leq \zeta_i \\ &< \zeta_i + 1 - (\beta - \alpha) \\ &= \zeta_{i-1} + 1 \\ &= \xi_{i-1}. \end{aligned}$$

Hence $j > i - 1$, i.e. $i \leq j$, and therefore $\zeta_i < \xi_j$.

Now for any $0 \leq i \leq n$, $\zeta_i \leq i\alpha^{-1} < (i+1)\alpha^{-1} < \xi_{i+1}$, and also we have

$$T\left([i\alpha^{-1}, (i+1)\alpha^{-1}]\right) = [\zeta_i, \xi_{i+1}).$$

Thus regardless of how i and j are chosen, as long as $0 \leq i \leq j \leq n$ we would have $\overline{T(\Psi)} \supseteq \Psi$. Now, as $\zeta_i < \zeta_{i+1}$ and $\xi_i < \xi_{i+1}$ for any i, we only have to prove that for the i and j described in the statement, $T\left([\zeta_i, i\alpha^{-1}, \xi_j]\right) \subseteq \Psi$ and $T\left([j\alpha^{-1}, \xi_j]\right) \subseteq \Psi$.

Note that

$$\sup_{x < i\alpha^{-1}} T(x) = \xi_i \leq \xi_j.$$

Also, as $\lambda_{i-1} \leq \zeta_i \leq i\alpha^{-1}$, if one takes $\zeta_i \leq \tilde{x} < i\alpha^{-1}$ then $T(\zeta_i) \leq T(\tilde{x}) < \xi_i$. Observe that $T(x) - x$ is continuous and increasing on $(\zeta_i, i\alpha^{-1})$ interval, and because $\lambda_{i-1} \leq \tilde{x}$ and λ_{i-1} is a fixed point, one has that $T(\zeta_i) > \zeta_i$. Therefore if $\zeta_i \leq \tilde{x} \leq i\alpha^{-1}$ then $T(\tilde{x}) \in \Psi$. An analogous argument proves that $T\left([j\alpha^{-1}, \xi_j]\right) \subseteq \Psi$. Note that by definition, Ψ is a closed set and we have $T(\Psi) \subseteq \Psi \subseteq \overline{T(\Psi)}$, so $\overline{T(\Psi)} = \Psi$. ∎

Of the invariant sets described by Proposition 8, the smallest of them is $[\zeta_m, \xi_n]$ where $m = \max\{i : \lambda_{i-1} \leq \zeta_i\}$ and $n = \min\{i : \xi_i \leq \lambda_i\}$. We shall denote it by $\Omega_{\beta\alpha}$ or Ω where the choice of α and β is clear from the context.

4.2. **Li-Yorke Theorem and Ergodicity of T for $K = 1$.** As it has already been proved, given α and β with $\beta > 1$, $\alpha \leq \beta$ and $\beta - \alpha < 1$ we have $\overline{T(\Omega_{\beta\alpha})} = \Omega_{\beta\alpha}$. Note that T is a piecewise monotone C^∞ function. Furthermore, let Ω^* be the set where both T and dT/dx are continuous. Then

$$\inf_{x\in\Omega^*} \left|\frac{d}{dx}T(x)\right| > 1$$

In [7], Lasota and Yorke proved that under these conditions there exist at least one non-negative function f of bounded variation such that the measure μ with $d\mu = fdm$ (where m is the Lebesgue measure) is invariant under T, in the sense that

$$\mu(E) = \int_E fdm = \int_{T^{-1}(E)} fdm = \mu\left(T^{-1}(E)\right).$$

In a more general setting, let $\tau : I \to I$ be piecewise twice continuously differentiable. Denote I^* the set of points where $d\tau/dx$ exists, and assume that

$$\inf_{x\in I^*} \left|\frac{d}{dx}\tau(x)\right| > 1. \tag{4.2}$$

We will refer to the points of $I - I^* = \{x_1, \ldots, x_k\}$ as the *points of discontinuity*. For $x \in I$, let $\Lambda(x)$ be the set of limit points of $\tau^n(x)$, that is

$$\Lambda(x) = \bigcap_{N=1}^{\infty} \overline{\{\tau^n(x)\}_{n=N}^{\infty}}. \tag{4.3}$$

An important property of this set is that it is fixed under τ, i.e. $\tau(\Lambda(x)) = \Lambda(x)$. Let $\mathcal{F}$ be the set of $f \in L^1(I)$, such that f is an invariant density under τ. In [8], Li and Yorke proved the following theorem.

Theorem 9. *Let $\tau : I \to I$ be a piecewise continuous and twice continuous differentiable interval map satisfying (4.2). Then, there exists a finite collection of sets $L_1, L_2, \ldots, L_n$ and a set of functions $\{f_1, f_2, \ldots, f_n\}$ such that*

1. *Each L_i is a finite union of closed intervals,*
2. *$L_i \cap L_j$ contains at most a finite number of points when $i \neq j$;*
3. *each L_i contains at least one point of discontinuity x_j, $1 \leq j \leq k$ on its interior; hence $n \leq k$;*
4. *$f_i(x) = 0$ for $x \notin L_i$ and $f(x) > 0$ for almost all x in L_i;*
5. *$\int_{L_i} f_i(x)dx = 1$ for $1 \leq i \leq n$;*
6. *if $g \in \mathcal{F}$ satisfy both (4) and (5), then $g = f_i$ almost everywhere;*
7. *every $f \in \mathcal{F}$ can be written as $f = \sum_{i=1}^{n} a_i f_i$ for suitably chosen $\{a_i\}_{i=1}^n$;*
8. *for almost every $x \in I$ there is an index i such that $\Lambda(x) = L_i$.*

Now assume that $1 < \beta < 2$ and $\beta(\beta-1) < \alpha < \beta$. Then $T(x) = \beta x - \lfloor \alpha x \rfloor$ generates a 1-bit quantization for every $x \in [0, 1)$. Now, by Proposition 8 we have $\Omega_{\beta\alpha} = [\alpha^{-1}\beta - 1, \alpha^{-1}\beta]$. This interval contains a unique point of discontinuity for T and T'. By Theorem 9, up to

normalization, there exists a unique non-negative function $f \in L^1$ such that the measure $d\mu = f dm$ is invariant under T. As this measure is unique, T is ergodic with respect to μ.

Indeed, the density of this function can be given in a closed form. In [9], Parry proved that if τ is a linear transformation $\pmod 1$, (i.e. $\tau(x) = bx + a \pmod 1$ for real numbers a and b), then the function

$$h(x) = \sum_{x<\tau^n(1)} \frac{1}{\beta^n} - \sum_{x<\tau^n(0)} \frac{1}{\beta^n},$$

where $\tau^0(x) = x$ by definition, is the density of an invariant measure of τ (potentially signed). Now if α and β are the parameters of a 1-bit $\beta\alpha$-encoder, we can define $b = \beta$, $a = (\beta-1)(\beta-\alpha)\alpha^{-1}$, and $f(x) = x - (\beta-\alpha)\alpha^{-1}$. Then $T(x) = f^{-1}(\tau(f(x)))$. By Parry's theorem, the function

$$g(x) = \sum_{x<T^n(\beta\alpha^{-1})} \frac{1}{\beta^n} - \sum_{x<T^n((\beta-\alpha)\alpha^{-1})} \frac{1}{\beta^n}$$

is the density of an absolutely continuous signed measure on $\Omega_{\beta\alpha}$, and by Li-Yorke's Theorem, such a measure is necessarily unique up to a re-scaling factor. Therefore, the density of the unique normalized invariant measure under T is

$$f(x) = \frac{1}{F}\left(\sum_{x<T^n(\beta\alpha^{-1})} \frac{1}{\beta^n} - \sum_{x<T^n((\beta-\alpha)\alpha^{-1})} \frac{1}{\beta^n}\right),$$

where F is a normalizing factor.

4.3. **Invariant Sets of T for $K > 1$.** A natural question at this point is: If $K > 1$, is there a unique (up to scaling) measure μ, that is absolutely continuous with respect to the Lebesgue measure and ergodic with respect to T? The answer in general is no.

Consider the $\beta\alpha$-encoder with $\alpha = 3/4$ and $\beta = 3/2$. In this case it is easy to show that $K = 2$ and $\Omega_{\beta\alpha} = [1, 3]$. One may check that T has two different invariant sets, namely $[1, 2]$ and $[2, 3]$. Therefore, by the theorem of Li and Yorke, there is a measure invariant under T for each of these intervals, each one independent of the other.

Notice that in this case, λ_1, ξ_1 and ζ_2, as defined in (4.1), are all equal. Our simulations suggesst that if for every index i the three numbers λ_i, ξ_i and ζ_{i+1} are distinct, then the system is indeed ergodic. One may conjecture that in this case there is a unique invariant measure which is ergodic. We leave this as an open question.

References

[1] I. Daubechies, R. DeVore, S. Güntürk and V. Vaishampayan, Beta Expansions: A New Aproach to Digitally Corrected AD Conversion, *IEEE International Symposium on Circuits and Systems*, **2** (2002), 784–787.

[2] I. Daubechies, R. DeVore, S. Güntürk and V. Vaishampayan, A/D Conversion with Imperfect Quantizers, *IEEE Transactions on Information Theory*, **52** (2006), 874–885.

[3] I. Daubechies, S. Güntürk, Y. Wang and Ö. Yılmaz, The golden ratio encoder, preprint.

[4] I. Daubechies, Ö. Yılmaz, Robust and Practical Analog-to-Digital Conversion With Exponential Precision, *IEEE Transactions on Information Theory*, **52** (2006), 3533-3545.
[5] A. Gersho and R. Gray, *Vector Quantization and Singal Compression*, Kluwer Academic Publishers, Boston, 1991.
[6] A. Lasota and M. C. Mackey, *Chaos, Fractals and Noise: Stochastic Aspects of Dynamics*, Springer-Verlag, New York, 1994.
[7] A. Lasota and J. A. Yorke, On the existence of invariant Measures for Piecewise Monotonic Transformations, *Transactions of the American Mathematical Society*, **186** (1973), 481–486.
[8] T. Li and J. A. Yorke, Ergodic Transformations from an Interval Into Itself, *Transactions of the American Mathematical Society*, **235**, (1978), pp. 183–192.
[9] W. Parry, Representations for Real Numbers, *Acta Mathematica Hungarica*, **15** (1964), 95–105.
[10] A Rényi, Representations for Real Numbers and their Ergodic Properties, *Acta Mathematica Hungarica*, **8** (1957), 477–493.
[11] N. Sidorov, Almost every number has a continuum of β-espansions, *A*merican Mathematical Monthly, **110** (2003), 838-842.
[12] M. Tsujii, Absolutely continuous invariant measures for expanding piecewise linear maps, *Inventiones Mathematicae*, **143** (2001), 349–373.

Department of Mathematics, Texas A&M University, College Station, Texas 77843, USA.

E-mail address: djimenez@tamu.edu

Department of Mathematics, Michigan State University, East Lansing, MI 48824, USA.

E-mail address: ywang@math.msu.edu